PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 | A2

(11) International Publication Number: WO 97/38376

(43) International Publication Date: 16 October 1997 (16.10.97)

(21) International Application Number: PCT/IB97/00748

(22) International Filing Date: 4 April 1997 (04.04.97)

(30) Priority Data:

60/014,815	4 April 1996 (04.04.96)	US
08/660,478	7 June 1996 (07.06.96)	US

(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) Applicant: FLAIR TECHNOLOGIES, LTD. [IL/IL]; 1 Corazin Street, 53583 Givatayim (IL).

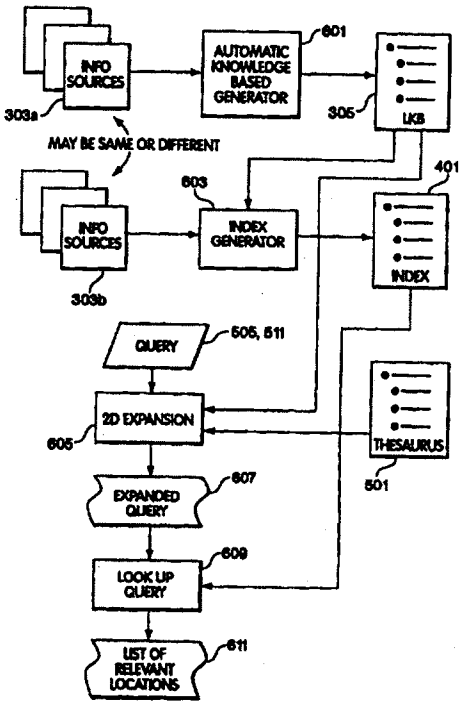
(72) Inventors: LEVI, Yuval; 9 Reuven Street, 93510 Jerusalem (IL). MARGULIS, Haim; 11 Nahal Snir Street, 77717 Ashdod (IL). ARAD, Iris; 10 Arazim Street, 96182 Jerusalem (IL).

(74) Agent: GADOR, Deborah; Seligsohn & Gabrieli, P.O. Box 1426, 61013 Tel Aviv (IL).

(54) Title: A SYSTEM, SOFTWARE AND METHOD FOR LOCATING INFORMATION IN A COLLECTION OF TEXT-BASED INFORMATION SOURCES

(57) Abstract

A system for processing information contained in a collection of text-based information sources employs associative and linguistic expansion of input words in which associative expansion is first performed, followed by simultaneous linguistic expansion in accordance with related morphological and phonetic rules. The system automatically generates and updates a linguistic knowledge base for each language to be processed by analyzing a large body of text in each language. The system also automatically indexes the collection of text-based information sources to be searched. A method is provided to expand a word or term in a supported language using a two-dimensional (2D) expansion matrix providing great flexibility, high accuracy and low noise output. The 2D expansion matrix includes an associative dimension that utilizes thesauri, databses of saved queries and other associated information sources, in which words are related to other words by meaning and relations, and a linguistic dimension which utilizes recognition-grammars, in which words are related to other words by combined rules for morphological and phonetic variation.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

A SYSTEM, SOFTWARE AND METHOD FOR LOCATING INFORMATION IN A COLLECTION OF TEXT-BASED INFORMATION SOURCES

BACKGROUND

1. Field of the Invention

The present invention relates generally to the field of information retrieval. More particularly, the present invention relates to information management systems and computational linguistic systems for finding information related to a user-input query, in a collection of text-based information sources.

2. Discussion of Related Art

In the Information Age, the ability to manage enormous volumes of information efficiently and find needed information quickly has become a driving force in all human endeavors. Early in the development of information management systems, the capability to process large volumes of free-form text documents and other text-based information sources was severely limited. Therefore information specialists developed various types of database management systems and searching systems based on strictly controlling how data may be received, stored, and referred to. However, as the volume and nature of the information which must be handled by such systems has expanded, conventional database management systems have been unable to keep pace.

In conventional database management systems, data is stored in a strictly structured environment. Such systems may be based upon tables of records or spreadsheet models, for example. Such systems may be flat or may be relational with respect to how records in the database are associated with each other. However, conventional database management systems generally require structured records in which one or more fields may be searchable, i.e. are key fields. Furthermore, it is desirable that such key fields use terms, e.g. numbering systems, labels, etc., in a consistent manner which facilitates searching with known query values, i.e. combinations of numbers, labels, etc.

In order to locate information within general text-based information sources, so-called full-text searching has developed. Full-text searching of a collection of text-based information sources, such as English-language documents stored in a computer system, permits a user to write a query containing terms known to be used in relevant documents. The collection of documents is first fully indexed and the words of the documents in the index are compared with

the query terms. In the simplest form of this type of system, an exact match between a query term and an index entry must be found in order to identify a relevant document. Spelling errors, word variants, etc. will tend to prevent finding all relevant documents. A technique called wild-carding may be used to partially alleviate this problem, but many irrelevant documents, referred to as "noise," often turn up when wild-carding is used. An example of the use of wild-carding is where a user query term includes only what the user has identified to be a word base of a relevant term, such as "comput*" for the concept of "compute," "computer," "computing," "computation," etc., where "*" indicates the portion of the term which has been left out.

Modern, conventional, full-text searching systems have been developed which have a much higher level of sophistication. For example, Pinkas, G., *Natural Language Full-Text Retrieval System*, Master's Thesis, University of Jerusalem, 1985, discloses a system which automatically expands a user's query to include additional relevant terms in a manner more noise-free than simple wild-carding. The Pinkas system: (1) receives a user query composed of query-words and boolean operators; (2) expands the query linguistically, i.e. by referring to a pre-processed database of morphological and phonetic information; (3) expands the query associatively, i.e. by referring to a database of associated sub-queries; and (4) merges the results of steps 2 and 3 above. Morphological expansion draws in the infix variations of the query terms, while phonetic expansion brings in terms that may be generated by misspelled vowels (e.g. recieve → receive). Associative expansion draws into the query related terms as pre-defined by the user in the form of sub-queries being associated with a specific query-word (e.g. to associate the acronym "USA" with its full wording, one creates an association between the word "USA" and a query applying a boolean "and" operation to the following 4 words: "United", "States", "of", "America", restricted to a proximity of 1 word distance. Thus a comprehensive expanded query is generated to cover the many different words that may conceptually be related to user's original query. Some variation in the level of morphological expansion and the level of phonetic expansion to be performed is available to the user by selection of expansion parameters.

However, this process of morphological and phonetic expansion suffers from many inefficiencies: it fails to recognize the fundamental differences between different "word-bases" such as morphological stems and phonemes, therefore it misses many relevant linguistic permutations affected by both mechanisms, and at the same time it generates a large amount of noise, i.e. false-positives, due to the combinatorial effect of combining both mechanisms. Moreover, this process also is fairly limited to recognizing and expanding single words, and even

then the interaction between the associative expansion and the linguistic expansion is fairly limited to a trivial merge of both results, having not shared a conceptual foundation that allows a mutual feedback (e.g. the query-word "airplane" expands to ("airplane" or "airplanes" or "aircraft") but not to "aircrafts".

In conventional systems, query expansion depended upon a set of linguistic rules which were developed by an expert in the language to be processed. The set of linguistic rules was both extensive and relatively inflexible, since as many characteristics of the input language as possible had to be accounted for before processing any text-based information sources. Development of the linguistic rules for each language to be processed was a very labor-intensive and time-consuming task.

Finally, conventional systems are known which require manual indexing of text sources, as well as which index text sources automatically. Conventional indexes simply map a word found in the text sources to a location at which the word is found. Manual full-text indexing is extremely time-consuming and error-prone. Keyword indexing is subjective and also somewhat error-prone.

SUMMARY OF THE INVENTION

Therefore, it is a general aim of the present invention to solve the problems noted above with respect to the prior art. Aspects of the present invention solving the problems of the prior art include at least a system, software and a method for processing information contained in a collection of text-based information sources.

The system may include a computer or data processor and software structured as one or more software modules, units or functions which when executed in a specified order by the computer or data processor perform the desired information processing task. One or more software modules, units or functions may be made available in conventional manner as either compile-time or run-time library entries which may be referred to by a software program which is written in a manner to be aware of such a library. The present invention further provides a method to process query-concepts and transform them to an expanded/improved query using an expansion matrix providing great flexibility, high accuracy and low noise output.

According to one aspect of the invention, there may be provided a text-based information processing system, comprising an automatic linguistic knowledge base generator having an input receiving a collection of text-based information sources and which produces a linguistic

knowledge base; an index generator having inputs receiving a collection of text-based information sources and the linguistic knowledge base and which produces an index of the received text-based information and further which updates the linguistic knowledge base to reflect the inputs to the index generator and maintain correlation between the index and the linguistic knowledge base; a query processor having inputs receiving a query composed by an operator, the linguistic knowledge base, the index and a thesaurus and which produces a list of locations in the collection of text-based information sources relevant to the query. The text-based information processing system may be subject to numerous modifications and variations. For example, the automatic linguistic knowledge base generator, the automatic index generator and the query processor may be embodied in various ways.

In accordance with another aspect of the invention, in a text-based information processing system, an automatic linguistic knowledge base generator may comprise a parser, receiving an input stream of terms and producing individual terms; a language recognizer connected to receive the individual terms from the parser and which produces an output indicative of a language to which each individual term belongs; a normalizer connected to receive the individual terms and further to receive linguistic rules for the language indicated by the output of the language recognizer and producing normalized terms; and a linguistic expander connected to receive the legal individual terms and producing entries stored in the linguistic knowledge base.

In accordance with yet another aspect of the invention, in a text-based information processing system, an automatic indexer may comprise a parser, receiving an input stream of terms and producing individual terms; a language recognizer connected to receive the individual terms from the parser and which produces an output indicative of a language to which each individual term belongs; a normalizer connected to receive the individual terms and further to receive linguistic rules for the language indicated by the output of the language recognizer and producing normalized terms; and an index entry generator connected to receive the legal individual terms and producing entries stored in the index when the terms have not previously been indexed and modifying an existing index entry when the terms have previously been indexed.

Finally, in accordance with yet another aspect of the invention, in a text-based information processing system, an expansion unit for expanding terms in a language may comprise an associative expander having an input receiving a term and having an output representing the term and at least one associated term found by the associated expander making

reference to a thesaurus; and a linguistic expander having an input connected to the output of the associative expander and having an output representing the input of the linguistic expander and at least one term linguistically related to the input of the linguistic expander and found by reference to a linguistic knowledge base for the language.

The normalizers recited above may be constructed of two units. The first normalizer unit may be connected to receive the individual terms and the linguistic rules and producing terms from which illegal characters have been removed; and the second normalizer unit may then be connected to receive the terms from which illegal characters have been removed and the linguistic rules and which produces normalized terms including word stems found by applying the linguistic rules to the terms from which illegal characters have been removed.

The present invention will be better understood by reading the Detailed Description of at least one illustrative embodiment of the invention, in connection with the attached drawing.

BRIEF DESCRIPTION OF THE DRAWINGS

In the drawings, in which like reference designations indicate like elements,

Fig. 1 is a schematic block diagram of a computer or data processing system on which the present invention may be practiced;

Fig. 2 is a schematic block diagram of the memory of Fig. 1;

Fig. 3 is a flow chart of automatic linguistic knowledge base generation;

Fig. 4 is a flow chart of automatic index generation;

Fig. 5 is a flow chart of query expansion; and

Fig. 6 is a flow chart of an information retrieval system including the features illustrated in Figs. 3-5.

DETAILED DESCRIPTION

In order to better understand the following detailed description, reference should be made to the following definitions. In this discussion a "language" is considered to be any organized system of tokens, which have symbolic meaning. For convenience, the tokens are referred to hereinafter as "words" or "terms," since the most common types of languages dealt with by text-based information systems are natural human languages composed of words or combinations of words, i.e. terms, which are understood to have specific meanings by humans. Thus, the terms "words" and "terms" are intended to encompass word phrases in those instances where a word

phrase may in fact be a token having a meaning independent of its sub-units, keywords in those instances where a word or word-phrase has a specific context/importance, and artificial words such as acronyms and short-cuts. A "word base" is the base portion of a word which remains after removing all prefixes and suffixes of the word which modify the meaning or part of speech of the word root appropriately for the context in which the word may be used. The term "thesaurus" as used herein refers to a database of terms, words and/or word bases, in which each term, word or word base is associated in the database with other terms, words and word bases having a defined relationship such as morphological proximity, phonetic similarity, similar meaning (synonyms), nearly opposite meaning (antonyms), broader meaning, narrower meaning, related term in a specific context, etc. The database may be navigated or searched on the basis of the terms, words and/or word bases stored therein.

Languages considered here have known linguistic rules for the morphological and phonetic variations which words may undergo. For example, the morphological rules of a language may define how a plural is formed from a singular noun, by changing the shape of the word, i.e. adding a final "s" in English, while the phonetic rules may represent the common variations in spelling resulting from user spelling errors. A table, file or database may be used by a software program to hold a list of such linguistic rules.

Generally, languages also include words which do not follow the linguistic rules of the language. For example, the English language morphological rule for generating the past tense of a verb does not apply to the verb "to go," which becomes "went" rather than the nonsensical "goed." Therefore, exceptions to the rules may be held by a software program in a table of exceptions, so that words which do not obey the rules may be handled as accurately as words which do obey the rules. In the context of the present invention, a "linguistic knowledge base" is developed by applying the linguistic rules and the one or more tables of exceptions or irregular forms to a large body of textual information to produce an efficient, adaptable and flexible representation of the variations of word bases which produce meaning, particularly in natural languages, but also in language generally. The "linguistic knowledge base" is a table, list or database of word bases and related words. Related words are those words which when analyzed under the linguistic rules for the language are determined to have the same word base.

The present invention is constructed in the context of computer systems and data processing systems. An overview of such systems is given in connection with the block diagram of Fig. 1. A computer system or data processing system generally includes a processor 101, a

memory 103, one or more input devices 105, and one or more output devices 107, all interconnected through an interconnection mechanism 109. Many variations of this basic plan are possible. For example, viable systems may lack input devices 105 and output devices 107, communicating entirely through interactions with the memory 103 by external devices (not shown). Also, distributed computer systems and data processing systems are contemplated as falling within this basic plan. The interconnection mechanism 109 may be an internal system bus of a personal computer or may be the Internet, through which a processor 101 interacts with a database stored on a remote memory 103. Other variations will be evident to those skilled in this art.

Memory 103 may be classified into two categories useful to this discussion, long term memory (also called non-volatile memory), and short term memory (also called volatile memory). These two types of memory are often both used in computer systems and data processing systems, as shown in Fig. 2. Volatile memory 201 such as integrated circuit random access memory (RAM) is often used in close physical proximity to the processor 101 because the technologies in which such volatile memory 201 is most readily realized produce fast access times, such as are desirable to support fast processors 101. Non-volatile memory 203 is often used to store massive quantities of data for longer periods of time because it can be more cheaply constructed than volatile memory of a similar capacity. Non-volatile memory 203 is often implemented as magnetic or optical disk or tape storage units, which provide a further advantage of data and software program interchange between different computer or data processing systems. As such, non-volatile memory 203 may be a software product disk on which are recorded signals representing instructions, which when executed by a processor 101 cause the computer or data processing system to perform a special purpose function. Software embodying aspects of the present invention may be recorded on such a non-volatile memory 203 for distribution by a manufacturer, for archival purposes, for access through a volatile memory 201 by a processor 101, etc.

In accordance with various aspects of the present invention, there may be constructed a system for searching through and locating information in a collection of text-based information sources. In accordance with various aspects of the invention, a linguistic knowledge base is first generated. Then, the collection of text-based information sources is indexed. A user next inputs a query defining the information sought. The query is expanded according to selected

associative and linguistic rules, using a thesaurus and the linguistic knowledge base. Finally, information is identified which matches the various expanded query terms.

The thesaurus, linguistic knowledge base and index may be stored in one or more computer files to which the system has access through memory 103.

The aspects of the invention connected with automatic generation of the linguistic knowledge base, automatic generation of the index and query expansion are next described in detail.

I. Automatic Generation of the Linguistic Knowledge Base

According to one aspect of the invention, software as shown in Fig. 3 is provided which when executed by a suitable data processing system will automatically generate the linguistic knowledge base from an input body of text based information sources. For example, according to this aspect of the invention, a collection of English language documents may be processed to generate an English language linguistic knowledge base.

A small set of linguistic rules 301, including a list of exceptions 302 to the linguistic rules for a language, e.g. English, is first generated by statistical analysis of a large body of text based information. This small set of rules includes:

- a list of irregular words and word bases in the language, i.e. the list of exceptions noted above;
- a word normalization table specifying legal characters in the language, i.e. the alphabet of the language, and legal character positions in the language, e.g. special rules concerning characters which can only appear at specific locations within a word;
- a prefix and suffix list specifying legal prefixes and legal suffixes in the language; and
- letter-to-sound rules for both ordinary words and proper names in the language.

This set of rules 301, including the list of exceptions 302, is then used to analyze a body of text based information sources 303, to generate a linguistic knowledge base 305 specifically adapted from the body of text based information sources 303. The body of sources 303 may be selected to be sources from a particular field of endeavor in which future queries are expected to be made, for example. This will result in a linguistic knowledge base better able to cope with the specifics of that particular field of endeavor. The body of sources 303 from which the linguistic knowledge base 305 is derived may not be the same body of sources which is ultimately to be searched. However, automatically generating the linguistic knowledge base 305 from the body

of sources to be searched has the advantage that the linguistic knowledge base 305 so produced is particularly well adapted to the body of sources to be searched.

Automatic generation of the linguistic knowledge base 305 proceeds as follows. The body of text based information sources 303 forms an input stream of text 304 to the system. This input stream 304 is first parsed into words and terms 307 in accordance with either fixed word recognition rules or word recognition rules specific to one or more languages. The language of each of the words parsed from the input stream is then recognized 309. Once the language of a word has been recognized the word may be normalized 311 according to the linguistic rules 301 for the language. Irregular words may also be recognized at this point, since known irregular words are already in the list of irregular words 302 and hence need no further processing. The system may also identify as potential new irregular words, those words meeting some rule-based criteria. Those previously unknown irregular words may be identified to a human operator for a determination of whether they should be added to the list of irregular words. Regular words are linguistically expanded 313 before being added to the linguistic knowledge base 305 such that word bases are stored in the linguistic knowledge base 305 along with a list of related words from the body of sources 303. Linguistic expansion 313 is discussed in greater detail below.

The step of parsing 307 the input stream 304 into sentences and words takes place according to the following pseudo-code:

```
Load segmentation rules;
segment input stream into sentences and words using segmentation
rules;
     for each sentence
     {
          for each word
          {
               if language not explicitly specified, identify
                    word's language
               end if;
               normalize word;
               return word and word's coordinates;
               return rest of input stream;
          } next word;
     } next sentence.
```

Normalization is performed as follows. Normalization identifies and removes garbage characters from the words of the input stream 304.

```
For each character of an input word
{
     if the character is illegal
     {
          set normalization status according to the character;
     }
     else
     {
          translate the character to the internal alphabet;
          add the translated character to the output word;
     }
} next character.
```

Finally, new keys are added to the linguistic knowledge base 305 by the following procedure.

```
If language not explicitly specified
     identify the language of the input word;
for each recognition type
{
     analyze word according to recognition type and level;
     search for analysis results in key table;
     if result is found in key table
     {
          next recognition type;
     }
     else
     {
          insert key and word into table;
          if key has a legal sub-key
          {
               activate linguistic correction mechanism;
          }
     }
} next recognition type.
```

Two useful recognition types subject to analysis as indicated in the above pseudocode are morphological and phonological. The morphological analyzer of the described embodiment functions in accordance with the following procedure. The morphological analyzer receives a list of valid prefixes and suffixes in the language identified for the input word.

```
Start at end of word;
strip next substring from end;
for each substring of word
{ /* search for prefix*/
     if substring is found to be a prefix in the identified
          language
     {
          strip prefix - create initial stem;
          /* search for suffix */
          start at beginning of stem;
          strip next substring from beginning of stem;
          for each substring of stem
          {
               if substring is found to be a suffix
               {
                    strip suffix - create stem;
                    return stem;
               } endif;
          } next substring;
     } endif;
} next substring.
```

The phonetic analyzer converts each word into a phonological representation of the word on the basis of letter to sound rules. Words having similar or same phonological representations may be considered to be related by their phonetic morphology.

When the above processes have been completed for the body of text based information sources 303 initially presented, a linguistic knowledge base 305 for the languages of the text in the body of sources 303 will have been automatically generated. When new text based information sources are added to the system, they are also processed as above. Thus, new sources increase the knowledge and accuracy of the linguistic knowledge base 305 through the addition of new information to the linguistic knowledge base 305, as well as through the linguistic correction mechanism which corrects the contents of individual entries in the linguistic knowledge base 305 according to new information. The learning procedure which embodies the linguistic correction mechanism is as follows.

```
Open a new table entry for a new correct key;
get the list of words in the body of a previous key entry;
for each word
{
     re-analyze word;
     if analysis results match the new correct key
     {
          delete word from body of previous key entry;
          add word to body of new correct key entry;
     }
} next word;
if previous key entry is empty
     delete previous key entry.
```

When the system detects inconsistencies between the contents of the linguistic knowledge base 305 and a newly presented text source, the affected word base and list of related words may be automatically, or at the direction of a human operator, reanalyzed and updated in accordance with the newly presented information and the above procedure. Thus, the system constantly learns about each language processed and updates the affected linguistic knowledge bases.

II. Automatic Generation of the Index Correlated with the Linguistic Knowledge Base

In addition to the linguistic knowledge base 305, the retrieval system according to another aspect of the present invention shown in Fig. 4 automatically generates an index 401, whereby text based information may be found by reference to the index 401. Automatic generation of the index 401 is accompanied by updating of the linguistic knowledge base 305, so that the contents of the linguistic knowledge base 305 reflects the relevant terms contained in the body of text based information sources 303 and is thus correlated with the index 401. The index 401 simply relates words actually found in the body of text based information sources 303 to locations within the body of sources 303. It is preferred that the location be defined hierarchically. For example, the location may be represented hierarchically by a document number, section number, sentence number and position number. Other hierarchical location identification schemes may be used, as seen fit by those skilled in this art.

In accordance with a preferred embodiment of the invention, the index 401 is assisted by the linguistic knowledge base 305. The index 401 includes only words and terms actually occurring in the text-based information sources 303. The linguistic knowledge base 305 relates word bases derived from the words actually occurring in the text-based information sources 303 to lists of related words. During retrieval, which is explained below, the system retrieves an

entry from the linguistic knowledge base 305 which is then used to reference one or more index entries.

Automatic generation of the index 401 proceeds as follows. The body of text based information sources 303 forms an input stream of text 304 to the indexing subsystem. This input stream 304 is first parsed into words and terms 307 in accordance with the word recognition rules. The language of each of the words parsed from the input stream is then recognized 309. Once the language of a word has been recognized 309 the word may be normalized 311 according to the linguistic rules 301 for the language. An index entry is then generated 403 for each new normalized word. If the normalized word already has an entry in the index 401, then the location of the current occurrence of the word is added to the previous entry.

At substantially the same time as the above process, the linguistic knowledge base 305 is continuously kept correlated with new and modified entries produced in index 401. Each normalized word is reduced to its word base 405 in accordance with the linguistic rules of the language of the word. The word base and related word is then added to the linguistic knowledge base file 407, if not already present. The user may also specify that related words include various types of expansions of the word bases. If expansions are included, expansion of the word base is performed before storing the word base and related words in the linguistic knowledge base file 305. When indexing of a body of text based information sources 303 is complete, the linguistic knowledge base 305 is correlated with the index 401 and reflects the relevant terms contained in the body of text based information sources 303.

III. Query Expansion

Query expansion is performed in accordance with a third aspect of the invention, shown in Fig. 5. Since a query may contain more than one word or term, word recognition is first performed as above.

The words and terms identified by the word recognition task may further be normalized. That is, they may be converted to a base form, if desired. By making reference to a thesaurus and linguistic rules, spelling errors may be removed, different lexical forms of acronyms and short-cuts may be recognized, etc.

Each recognized word in the query may then be expanded using a 2D expansion matrix. The 2D expansion matrix is one way of defining the expansion space in which an input word may be represented. The dimensions of this space are associative and linguistic. The associative dimension is based upon the meaning of words/word-bases in the language to be processed. In

the described embodiment of the invention, the associative dimension is defined by one or more thesauri 501 relating words and terms to their synonyms, broader terms, narrower terms and other relations. Each thesaurus 501 includes a database of terms along with conceptually related terms. The thesaurus is searchable by term. Thus, each thesaurus entry contains an entry key which is a list of searchable terms. Each entry key has associated therewith one or more terms conceptually related to the entry key, such as synonyms, broader terms, narrower terms, associated terms, antonyms, etc. The inclusion of any one or more categories of association is optional. Furthermore, each entry term may optionally have associated therewith a conventional dictionary definition and usage guide, as well as a query string into which the entry key may be translated when required. Thus, the thesaurus is a list of entries, wherein each entry has a structure substantially as follows:

- KEYWORD: (in the form of natural language phrase or term) Used as an entry key.
- DESCRIPTION (optional): A description of keyword meaning and usage (as in encyclopedic dictionaries).
- QUERY: A complete query statement in an underlying full-text query language that the keyword is translated to when required (optional). (E.g., KEYWORD "USA" → QUERY "United AND States AND of AND America".) If a translation of the keyword to a complete query statement is not supplied explicitly, a default translation is applied to the keyword.
- RELATIONS
  - SYNONYMS: A list of keywords synonymous with the KEYWORD that comprise a concept or descriptor.
  - BROADER TERMS
  - NARROWER TERMS
  - ASSOCIATIONS
  - OTHER

All of these features may be used by an operator to determine whether associative expansion is having a desired effect.

The linguistic dimension of this expansion space is based upon the linguistic knowledge base 305 of the language to be processed. As discussed above, the linguistic knowledge base is built automatically from the actual corpora of the text-based information sources, independent of manually crafted linguistic dictionaries, and not being restricted to "legal" or "proper" words. In

this embodiment of the invention, linguistic expansion grammars of morphology and phonetics are supported.

The expansion task performs 2D expansion in substantially two main steps. First an associative expansion is performed at step 503, in which each input word of an input query 505 is expanded to a list of words 507 including words having defined relations to the input word. The associated words are found by making reference to the thesaurus 501. This expanded list of words 507 becomes the input on which linguistic expansion 509 is performed in both the morphological and phonetic dimensions, simultaneously. The morphological and phonetic expansion is controlled by making reference to the linguistic knowledge base 305. The linguistic expansion 509 may be controlled by expansion parameters 511 supplied by the user to include varying degrees of morphological expansion and phonetic expansion, ranging for both dimensions from no expansion in that dimension to full expansion in that dimension. By performing the morphological and phonetic expansions as a single, linguistic expansion step 509, expansion strategies for morphology and phonetics may be intelligently related. The relationships between the expansion dimensions are defined in the linguistic knowledge base 305 for the language. Thus, a rule for morphological expansion may define a morphological variation which changes depending upon the phonetic properties of the input word or the expanded result. As a result, less noise is generated in the expanded output because relating the morphological and phonetic dimensions as a single linguistic plane eliminates morphological variants which are phonetically unacceptable under the totality of the linguistic rules, and vice versa.

IV. A Complete Text Retrieval System

It can now be seen that using the software described above a retrieval system may be constructed as shown in Fig. 6, which can perform efficient and accurate location of information within a collection of text-based information sources. Briefly, such a system is given access to one or more collections of text-based information sources 303a and 303b. At least one group of text-based information sources 303a is supplied to automatic linguistic knowledge base generating software 601, which generates the linguistic knowledge base 305 as described above. Text-based information sources 303b are provided to an indexing subsystem 603 which creates an index 401 of words in the text-based information sources 303b, in which each entry in the index 401 defines a relationship between a word and the location of the word in the collection, as described above. It is preferred that the index 401 be generated using normalized words in one

or more languages for which the system has a thesaurus 501 and a linguistic knowledge base 305. The indexing subsystem 603 may include a module to recognize words having forms which conform to one of the languages supported by the system and may further include an appropriate normalizing module for each language supported by the system. Words are normalized in their language as discussed above, to reduce the number of anomalous entries appearing in the index 401. The system further receives a user query 505 in the form of one or more words expressive of the information sought by the user. The query words are expanded 605 using a 2D expansion matrix, as discussed above. The query is first associatively expanded to include words related to the original query words by reference to the thesaurus appropriate for the language of the query words. The associatively expanded query is then linguistically expanded in both the morphological and phonetic dimensions, simultaneously. The degree of expansion in each dimension is specified by the user, by parameters 511 supplied with the query. The degree of expansion may be specified by the user, for example, by attaching a checklist of expansion parameters 511 to each query term. Finally, the terms of the fully expanded query 607 are compared 609 with the entries in the index 401 to find relevant locations 611 within the collection of text-based information sources 303b.

Relevant locations 611 in the collection of text-based information sources 303b do not necessarily contain any of the original query terms. By the processing described above, the locations found 611 will contain one of the original query terms or a related term produced by the associative and linguistic expansion processes. The locations found 611 will not include many "noise" locations because the linguistic expansion process is performed as described above in a manner in which the morphological and phonetic linguistic rules are applied simultaneously in a synergistic manner that avoids the problem of applying a morphological rule to generate a phonetically nonsensical result or vice versa.

In a system such as described above, the text-based information sources may be text documents stored on a computer system. In this case, it may be convenient for the indexing system to hierarchically refer to locations by document number, section number, sentence number and position with the sentence. Furthermore, freely formatted text documents may be processed by the above-described system. There is no need to structure the documents a particular way or to manually produce classifications or keywords, as done in some prior art systems, because the present system indexes words and manipulates queries according to the rules of the language in which the words occur.

If it is desired that a phrase be treated as a single word or term in a particular language, then that phrase may be so defined as a conceptual entity in a thesaurus. In all other respects, the phrase so defined as a word is treated simply as a word in the language. However, it is unnecessary to declare a long list of accepted keywords because the process of indexing and query expansion generates accurate, relatively noise-free matches for user queries reasonably expressive of the information sought.

Having thus described at least one illustrative embodiment of the invention, various alterations, modifications, and improvements will readily occur to those skilled in the art. Such alterations, modifications and improvements are intended to be within the spirit and scope of the invention. Accordingly, the foregoing description is by way of example only and is not intended as limiting. The invention is limited only as defined in the following claims and the equivalents thereto.

CLAIMS

1. A text-based information processing system, comprising:

an automatic linguistic knowledge base generator having an input receiving a collection of text-based information sources and which produces a linguistic knowledge base;

an index generator having inputs receiving a collection of text-based information sources and the linguistic knowledge base and which produces an index of the received text-based information and further which updates the linguistic knowledge base to reflect the inputs to the index generator and maintain correlation between the index and the linguistic knowledge base; and

a query processor having inputs receiving a query composed by an operator, the linguistic knowledge base, the index and a thesaurus and which produces a list of locations in the collection of text-based information sources relevant to the query.

2. In a text-based information processing system, an automatic linguistic knowledge base generator, comprising:

a parser, receiving an input stream of terms and producing individual terms;

a language recognizer connected to receive the individual terms from the parser and which produces an output indicative of a language to which each individual term belongs;

a normalizer connected to receive the individual terms and further connected to receive linguistic rules for the language indicated by the output of the language recognizer and producing normalized terms; and

a linguistic expander connected to receive the normalized terms and producing entries stored in the linguistic knowledge base.

3. The system of claim 2, wherein the normalizer further comprises:

a first normalizer unit connected to receive the individual terms and the linguistic rules and producing terms from which illegal characters have been removed; and

a second normalizer unit connected to receive the terms from which illegal characters have been removed and the linguistic rules and which produces normalized terms including word stems found by applying the linguistic rules to the terms from which illegal characters have been removed.

4. In a text-based information processing system, an automatic indexer, comprising:

a parser, receiving an input stream of terms and producing individual terms;

a language recognizer connected to receive the individual terms from the parser and which produces an output indicative of a language to which each individual term belongs;

a normalizer connected to receive the individual terms and further connected to receive linguistic rules for the language indicated by the output of the language recognizer and producing normalized terms; and

an index generator having inputs receiving a collection of text-based information sources and the linguistic knowledge base and which produces an index of the received text-based information and further which updates the linguistic knowledge base to reflect the inputs to the index generator and maintain correlation between the index and the linguistic knowledge base.

5. The system of claim 4, wherein the normalizer further comprises:

a first normalizer unit connected to receive the individual terms and the linguistic rules and producing terms from which illegal characters have been removed; and

a second normalizer unit connected to receive the terms from which illegal characters have been removed and the linguistic rules and which produces normalized terms including word stems found by applying the linguistic rules to the terms from which illegal characters have been removed.

6. In a text-based information processing system, an expansion unit for expanding terms in a language, comprising:

an associative expander having an input receiving a term and having an output representing the term and at least one associated term found by the associated expander making reference to a thesaurus; and

a linguistic expander having an input connected to the output of the associative expander and having an output representing the input of the linguistic expander and at least one term linguistically related to the input of the linguistic

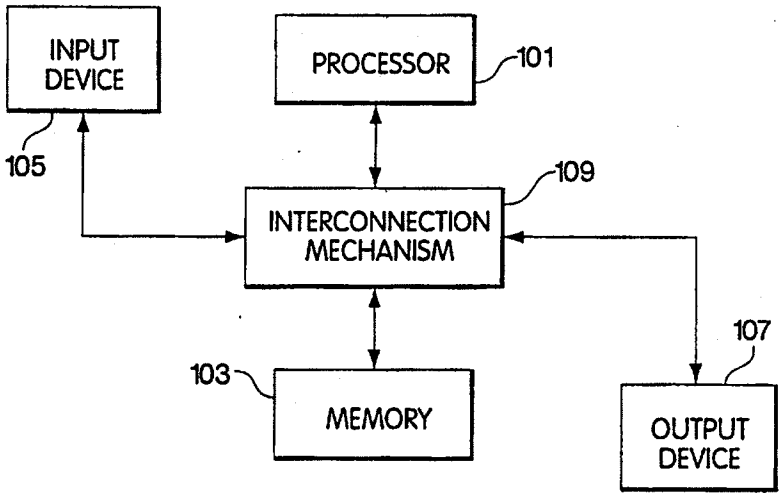

INPUT DEVICE
PROCESSOR
101
105
109
INTERCONNECTION MECHANISM
107
103
MEMORY
OUTPUT DEVICE

Fig. 1

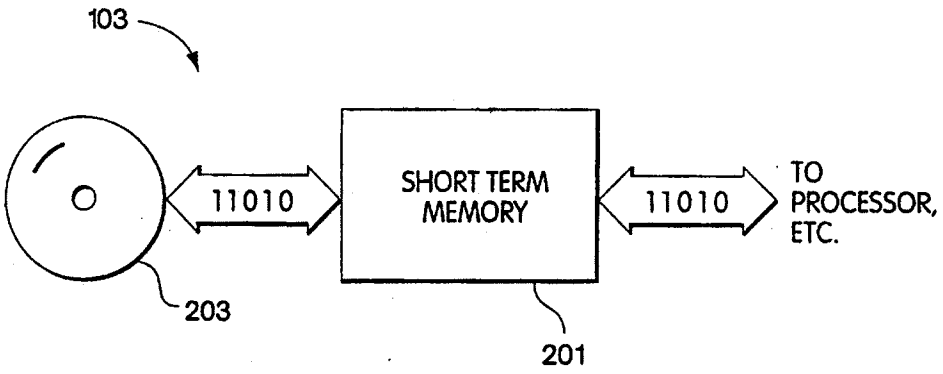

103
11010
SHORT TERM MEMORY
11010
TO PROCESSOR, ETC.
203
201

Fig. 2

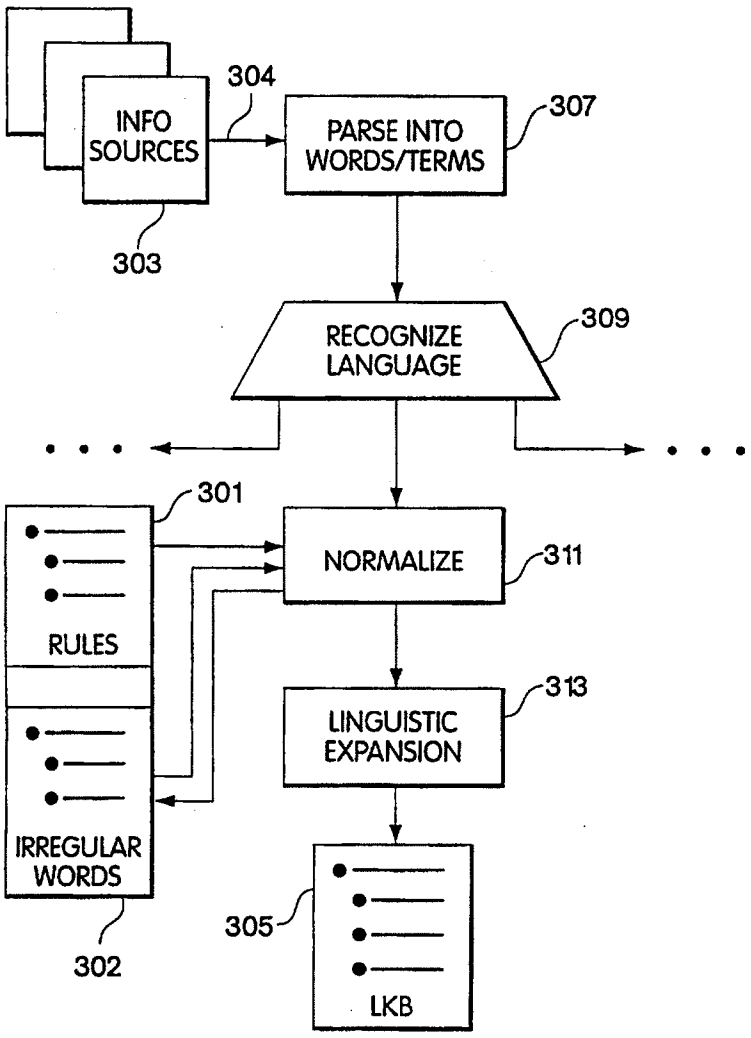

INFO SOURCES
304
303
PARSE INTO WORDS/TERMS
307
RECOGNIZE LANGUAGE
309
301
RULES
IRREGULAR WORDS
302
NORMALIZE
311
LINGUISTIC EXPANSION
313
305
LKB

Fig. 3

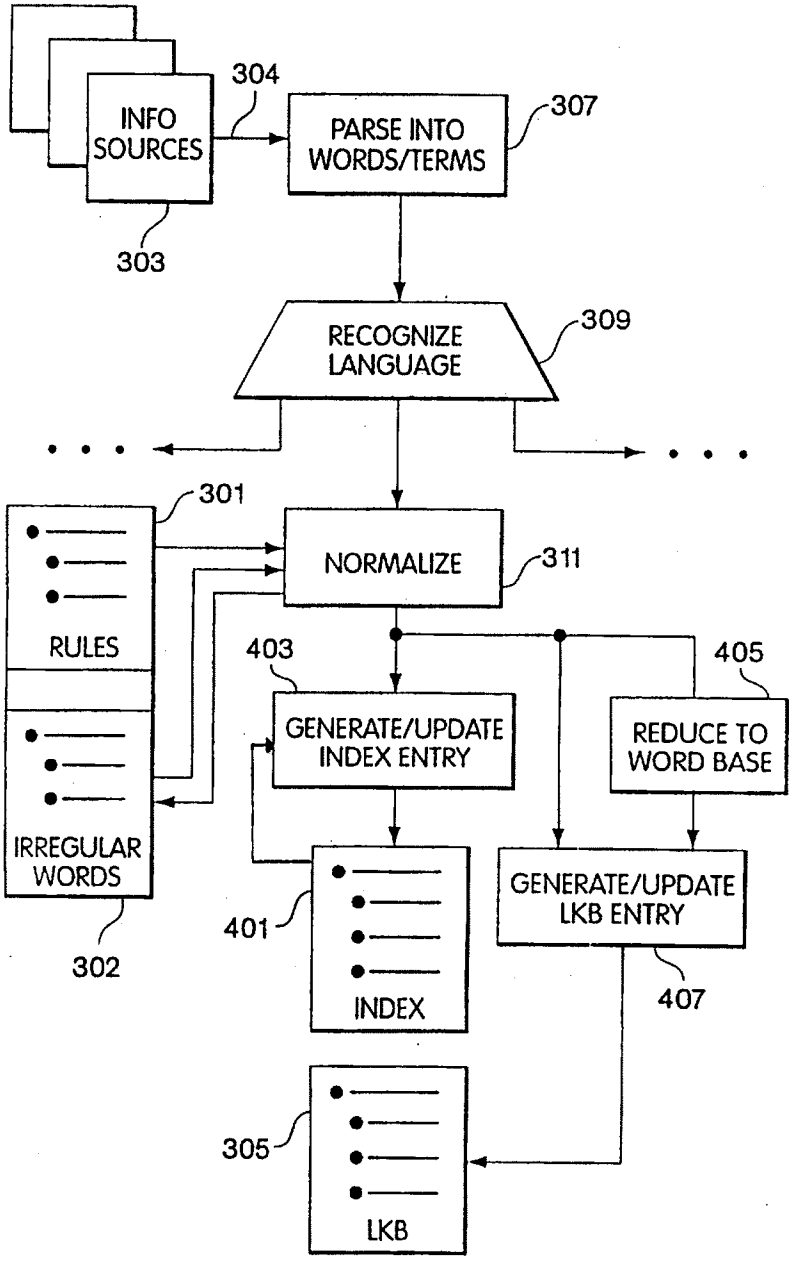

INFO SOURCES
304
303
PARSE INTO WORDS/TERMS
307
RECOGNIZE LANGUAGE
309
301
RULES
IRREGULAR WORDS
302
NORMALIZE
311
403
GENERATE/UPDATE INDEX ENTRY
401
INDEX
405
REDUCE TO WORD BASE
GENERATE/UPDATE LKB ENTRY
407
305
LKB

Fig. 4

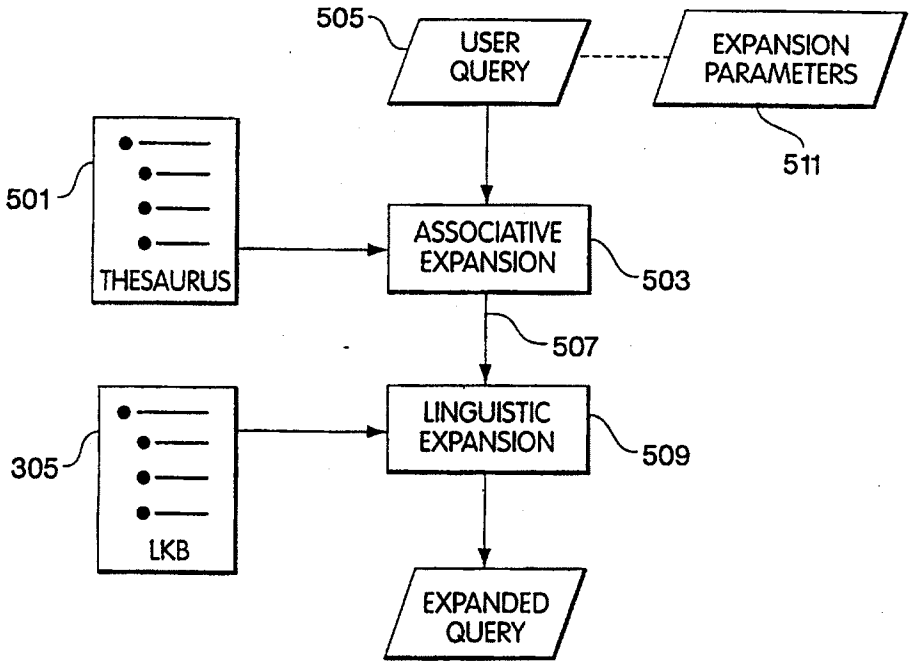

505
USER QUERY
EXPANSION PARAMETERS
511
501
THESAURUS
ASSOCIATIVE EXPANSION
503
507
305
LKB
LINGUISTIC EXPANSION
509
EXPANDED QUERY

Fig. 5

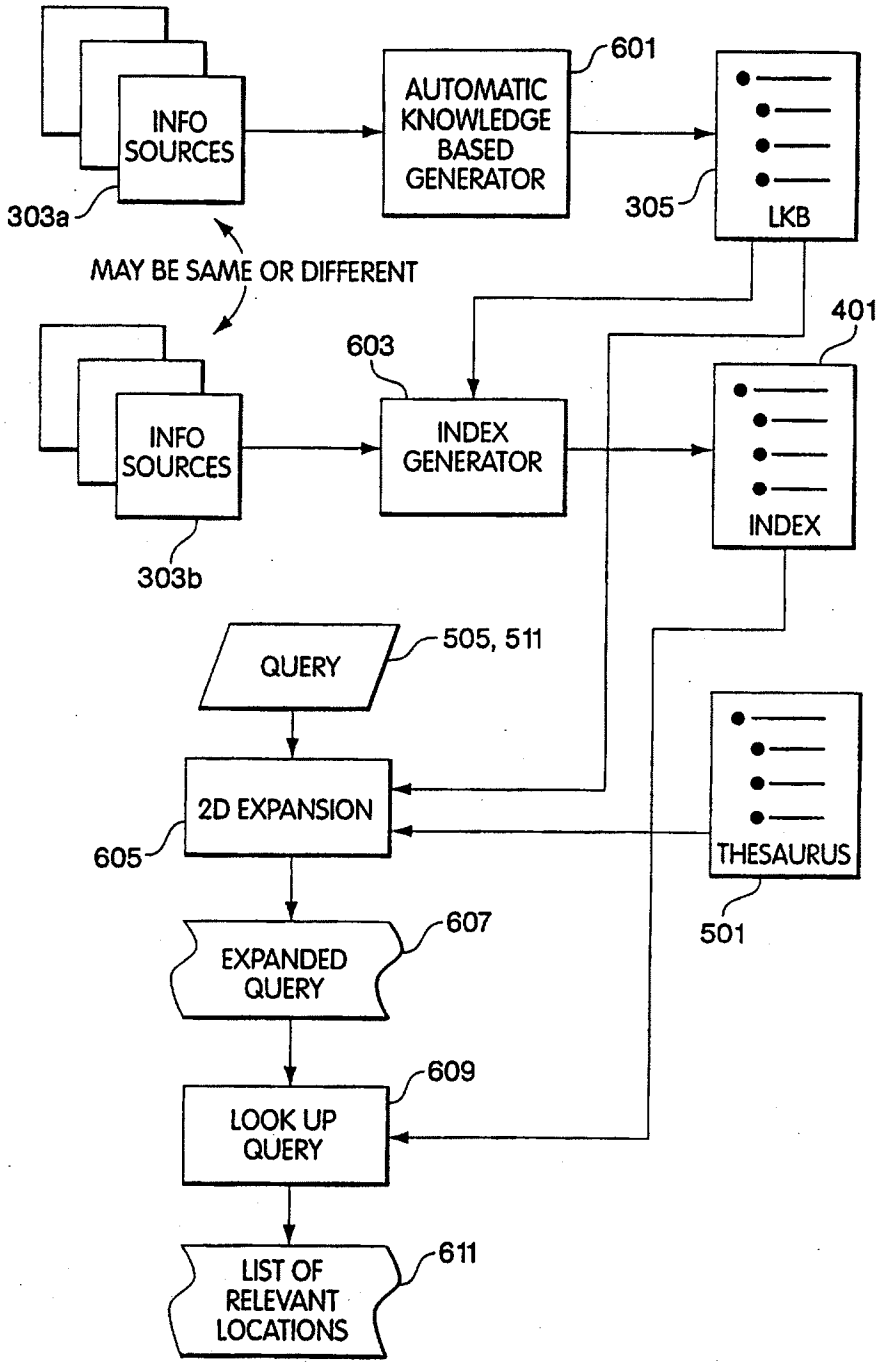

INFO SOURCES
303a
AUTOMATIC KNOWLEDGE BASED GENERATOR
601
LKB
305
MAY BE SAME OR DIFFERENT
INFO SOURCES
303b
603
INDEX GENERATOR
INDEX
401
QUERY
505, 511
2D EXPANSION
605
THESAURUS
501
EXPANDED QUERY
607
LOOK UP QUERY
609
LIST OF RELEVANT LOCATIONS
611

Fig. 6